## AMENDMENTS

### Amendments to the Specification

Please replace the paragraph on page 2 under the section heading "Related Applications" with the following amended paragraph:

This application is related to U.S. Patent Application Serial Number 09/721,042, filed on November 21, 2000, entitled "Methods and Computer Software Products for Predicting Nucleic Acid Hybridization Affinity"; U.S. Patent Application Serial Number 09/718,295, filed on November, 21, 2000, entitled "Methods and Computer Software Products for Selecting Nucleic Acid Probes"; U.S. Patent Application Serial Number 09/745,965, filed on 12/21/2000, entitled "Methods For Selecting Nucleic Acid Probes"; U.S. Patent Application Serial Number [[____]] 10/028,416, ~~attorney Docket No. 3439~~, filed on December 21, 2001, entitled "Method and Computer Software Product for Predicting Polyadenylation Sites" and U.S. Patent Application Serial Number [[____]] 10/027,682, ~~attorney docket number 3440,~~ filed on December 21, 2001, and currently abandoned. All the cited applications are incorporated herein by reference in their entireties for all purposes.

Please replace the paragraph beginning on page 4 and ending on page 5 with the following amended paragraph:

In some instances, methods for ~~triming~~ trimming a transcript sequence and detecting chimeric sequences are provided. The methods include aligning the transcript sequence to its ~~correponding~~ corresponding genomic sequence or sequences. Poorly aligned regions or regions which do not align can be treated as low quality while the

aligned portion can be treated as high quality. Furthermore, the low quality region can be removed, creating a "trimmed" version of the sequence containing only the high quality region(s). When mutually exclusive portions of the transcript sequence align to distant portions of the genome or align in a non-linear fashion, the transcript can be considered chimeric. Furthermore, two or more new sequences can be created based on the transcript alignments to the genome. "Distant" can refer to regions on different chromosomes, different strands of the same chromosome or regions sufficiently ~~appart~~ apart on the same chromosome and strand such that the distance is not likely to be an intron. Non-linear alignments occur when the order of the aligned regions in the genomic sequence is different than the order of the regions within the transcript. The quality of the genomic sequence can optionally be taken into account such that these annotations and actions occur only for genomic sequence of a specific quality (such as finished).

Please replace the paragraph beginning on page 5 and ending on page 6 with the following amended paragraph:

Methods are also provided for designing nucleic acid probe arrays using trimmed transcripts for probe selection. The methods include aligning a transcript sequence to its corresponding genomic sequence; ~~triming~~ trimming a side of the transcript sequence to obtain a trimmed transcript sequence if the side of the transcript sequence is poorly ~~align~~ aligned with the genomic sequence; and selecting probes targeting the trimmed transcript sequence or clusters including the trimmed transcript sequence.

Please replace the paragraph on page 10 with the following amended paragraph:

FIGURE 3 shows an exemplary computer network that is suitable for executing the computer software of the invention. A computer workstation 302 is connected with the application/data server(s) through a local area network (LAN) 301, such as an Ethernet 305. A printer 304 may be connected directly to the workstation or to the Ethernet 305. The LAN may be connected to a wide area network (WAN), such as the Internet 308, via a gateway server 307 which may also serve as a firewall between the WAN 308 and the LAN 305. In preferred embodiments, the workstation may communicate with outside data sources, such as the National Biotechnology Information Center, through the Internet. Various protocols, such as FTP and HTTP, may be used for data communication between the workstation and the outside data sources. Outside genetic data sources, such as the GenBank 310, are well known to those skilled in the art. An overview of GenBank and the National Center for Biotechnology information (NCBI) can be found in the web site of NCBI ~~("www.ncbi.nlm.nih.gov")~~.

Please replace the paragraph beginning on page 14 and ending on page 15 with the following amended paragraph:

Microarrays can be used in a variety of ways. A preferred microarray contains nucleic acids and is used to analyze nucleic acid samples. Typically, a nucleic acid sample is prepared from appropriate source and labeled with a signal moiety, such as a fluorescent label. The sample is hybridized with the array under appropriate conditions. The arrays are washed or otherwise processed to remove non-hybridized sample nucleic acids. The hybridization is then evaluated by detecting the distribution of the label on the chip. The distribution of label may be detected by scanning the arrays to determine

fluorescence intensity distribution. Typically, the hybridization of each probe is reflected by several pixel intensities. The raw intensity data may be stored in a gray scale pixel intensity file. The GATC™ Consortium has specified several file formats for storing array intensity data. The final software specification is available at ~~www.gatcconsortium.org~~ at the Consortium's website and is incorporated herein by reference in its entirety. The pixel intensity files are usually large. For example, a GATC™ compatible image file may be approximately 50 Mb if there are about 5 000 pixels on each of the horizontal and vertical axes and if a two byte integer is used for every pixel intensity. The pixels may be grouped into cells (see, GATC™ software specification). The probes in a cell are designed to have the same sequence (i.e., each cell is a probe area). A CEL file contains the statistics of a cell, e.g., the 75th percentile and standard deviation of intensities of pixels in a cell. The 50, 60, 70, 75 or 80th percentile of pixel intensity of a cell is often used as the intensity of the cell.

Please replace the paragraph on page 18 with the following amended paragraph:

A common way to assemble ESTs is by clustering. The goal of such a project is the construction of a gene index in which ESTs and full-length transcripts are partitioned into index classes (or clusters) such that they are placed in the same index class if and only if they represent the same gene. Projects related to EST clustering and assembly include UniGene from the National Center for Biotechnology Information; the TIGR Gene Index ~~(http://www.tigr.org/tdb/hgi/hgi.html)~~ from the Institute for Genomic Research; the Sequence Tag Alignment and Consensus Knowledgebase (STACK; maintained by the South African National Bioinformatics Institute) ~~http://ziggy.sanbi.ac.za/stack/stacksearch.htm~~); the Merck/Washington University Gene

Index; and the GenExpress project. All of these projects perform some type of cluster analysis in which sequence similarity is used to form the clusters. For an overview of EST clustering, see, Win Hide and Alan Christoffels, EST Clustering Tutorial, ISMB, 1999 (available at ~~www.sanbi.ac.za~~ South African National Bioinformatics Institute's website) and incorporated here by reference. It is worth noting that the gene indexing process typically incorporate information about EST and full length cDNA sequences.

Please replace the paragraph beginning on page 18 and ending on page 19 with the following amended paragraph:

In one aspect of the invention, transcript sequences (such as EST sequences, full length cDNA sequences, consensus or ~~examplar~~ exemplar sequences from sequence clusters, etc.) are aligned to the genomic sequence to obtain information or verify information about the transcriptome. For example, in a preferred embodiment, human transcript sequences are aligned to the human genome sequence to verify the validity of the orientation s using consensus splice sites, detect chimeric UniGene clusters, determine dbEST genomic ~~triming~~ trimming, etc. The alignment also provides information about transcribed locations in the genome.

Please replace the paragraph on page 19 with the following amended paragraph:

Alignment of the transcript ~~sequencies~~ sequences to the genomic sequence can be performed using for example, psLayout, a computer program available from University of California at Santa Cruz.

Please replace the paragraphs beginning on page 20 and ending on page 21 with the following amended paragraph:

3. Genomic annotation: At least two annotation alignments can be generated using the genomic position of the consensus. Any overlapping annotations (gene annotations and cDNA alignments) can be recorded as well as any splice consistent annotations. Alignments to the genome can be used to transfer annotations to the cluster. For instance, if a ~~partielar~~ particular disease loci is located in the same spot that the cluster is, then transfer that disease annotation to the cluster. Likewise, if an EST cluster that aligns to the same region as a gene prediction, the predicted gene annotations can be transferred to the cluster.

4. Cluster merging: Clusters with consensus and/or member sequences which overlap (e.g., within 1000 bases and optionally on the same strand) in genomic space, or clusters with consensus sequences and/or member sequences which are "evidence." As used herein, "evidence" means they overlap in genomic space or that they share exonic sequence for the same genomic annotation (such as a predicted gene ~~struction~~ structure) are considered for merging.